

Efficiently Learning the Preferences of People

Adriana Birlutiu · Perry Groot · Tom Heskes

Received: date / Accepted: date

Abstract This paper presents a framework for optimizing the preference learning process. In many real-world applications in which preference learning is involved the available training data is scarce and obtaining labeled training data is expensive. Luckily in many of the preference learning situations data is available from multiple subjects. We use the multi-task formalism to enhance the individual training data by making use of the preference information learned from other subjects. Furthermore, since obtaining labels is expensive, we optimally choose which data to ask a subject for labelling to obtain maximum of information about her/his preferences. This paradigm —called active learning— has hardly been studied in a multi-task formalism. We propose an alternative for the standard criteria in active learning which actively chooses queries by making use of the available preference data from other subjects. The advantage of this alternative is the reduced computation costs and reduced time subjects are involved. We validate empirically our approach on two real-world data sets involving the preferences of people: an audiological data set and an art image data set.

Keywords learning preferences, active learning, experimental design, multi-task learning, hierarchical modeling

A.Birlutiu
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
Tel.: +31 24 3652354
E-mail: A.Birlutiu@science.ru.nl

P.Groot
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
Tel.: +31 24 3652354
E-mail: Perry.Groot@science.ru.nl

T.Heskes
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
Tel.: +31 24 3652696
E-mail: T.Heskes@science.ru.nl

1 Introduction

Recently, there has been an interest in learning the preferences of people within artificial intelligence research (Doyle, 2004). Preference learning is a crucial aspect in modern applications such as decision support systems (Chajewska et al, 2000), recommender systems (Blythe, 2002; Blei et al, 2003), and personalized devices (Clyde et al, 1993; Heskes and de Vries, 2005). The preferences of a person are learned in order to make recommendations or take decisions on her/his behalf. For example, a movie recommender system needs to learn a person’s preferences about movies before making a recommendation about a movie she/he might prefer.

Since preference learning is a cumbersome process it is important to make it as efficient as possible in order to reduce the costs and time involved. In practice it is important that the time subjects are involved is minimized. The preferences should be learned accurately from a minimum number of training data.

In most of the situations in which preference learning is involved data is available from multiple subjects. Thus, even though individual data is scarce and difficult to obtain, we can optimize the learning of preferences of a new subject by making use of the available data from other subjects. Learning in this setting is well-known as multi-task or hierarchical learning and has been studied extensively in recent years in machine learning. By using the multi-task formalism, the preference data collected for other subjects can be gathered and used as prior information when learning the preference of a new subject. Furthermore, to deal with the fact that obtaining labeled data is expensive we can speed up learning by optimally choosing the examples to be queried. At each learning step we can decide which example gives the most information about the subject’s preferences. This paradigm, called active learning in the machine learning literature and related to sequential experimental design in statistics, has been studied extensively, but hardly in the multi-task setting.

The aim of this paper is to present an efficient framework for optimizing the preference learning process. This framework considers the combination between active learning and multi-task learning in the preference learning context. The contribution of this paper is a criterion for active learning designed for the multi-task setting. The advantages of this criterion are in its interpretation and the ease in computability.

1.1 Structure of this Article

The structure of this article is as follows.

In Section 2, we describe the learning framework. We consider learning from qualitative preference observations which can be modeled using the probabilistic choice models introduced in Section 2.1. Learning a utility function representing the preferences of a subject from this type of preference observations is described in Section 2.2. Learning the utility function in a multi-task setting by making use of the data available from other subjects is considered in Section 2.3.

In Section 3 we present several criteria for selecting the most informative experiments with respect to a subject’s preferences. After reviewing some of the standard criteria from experimental design, we propose an alternative criterion which makes use of the preference observations collected already from a community of subjects. We show that this alternative criterion is connected to the standard criteria from exper-

imental design. Furthermore, it has several advantages due to its interpretation and computational simplicity.

In Section 4 we demonstrate experimentally the usefulness of our framework on an audiological data set and a data set about people’s preferences for art images.

In Section 5 we conclude and discuss several directions for future research.

1.2 Related Work

In this section we review some studies from preference learning, multi-task learning, and active learning related to the work presented in this paper.

1.2.1 Preference Learning

The existing approaches to preference learning can be divided into two categories. The first category aims at obtaining a ranking of instances from a set of pairwise preferences by solving an augmented binary classification problem (Herbrich et al, 1998; Har-peled et al, 2002; Furnkranz and Hullermeier, 2003; Aioli and Sperduti, 2004). The second category uses regression to map instances to target valuations for direct ranking (Caruana et al, 1996; Crammer and Singer, 2001; Chu and Ghahramani, 2005b; Brochu et al, 2007). We focus on the latter and use a utility function in order to model the subject’s preferences. The utility function is learned in a Bayesian framework. By formulating the preference elicitation problem as a probabilistic Bayesian learning problem one can deal with inconsistencies in subjects responses as well as learn biases the subjects may have.

1.2.2 Multi-Task Learning

In this paper we implement multi-task learning in a Bayesian framework. The Bayesian approach to multi-task learning assumes the parameters of individual models to be drawn from the same prior distribution. The hierarchical prior can be a Gaussian distribution. In this case relatedness between tasks is captured as follows: the smaller the variance of the Gaussian the more related the tasks are. Examples of the Bayesian approach to multi-task learning are (Bakker and Heskes, 2003) where a mixture of Gaussians is used for the top of the hierarchy. This leads to clustering the tasks, one cluster for each Gaussian in the mixture. In (Yu et al, 2005; Birlutiu et al, 2009) a hierarchical Gaussian Process is derived with a normal-inverse Wishart distribution used at the top of the hierarchy.

Another approach to multi-task learning is regularization (Evgeniou et al, 2005; Argyriou et al, 2008). The basic idea in multi-task regularization is that task relationships are modeled through the assumption that the error terms (noise) for regressions estimated simultaneously are correlated.

1.2.3 Active Learning

Methods for active learning can be roughly divided into two categories: those with and without an explicitly defined objective function. Uncertainty sampling (Lewis and Gale, 1994), Query-by-Committee (Seung et al, 1992; Freund et al, 1997) and variants

thereof belong to the latter category. They are based on the idea of selecting the most uncertain data given the previously trained models. The methods with an explicit objective function are often motivated by the theory of experimental design (Fedorov, 1972; Chaloner and Verdinelli, 1995; Schein and Ungar, 2007; Lewi et al, 2009; Dror and Steinberg, 2008). The objective function then quantifies the expected gain of labeling a particular input, for example in terms of the expected reduction in the entropy of the model parameters (MacKay, 1992; Cohn et al, 1996). Schein and Ungar (2007) show that the methods from the second approach perform better but are computationally more expensive due to re-training the models for each candidate point.

A sub-trend is to improve the performance of the active learning methods by combining them with heuristics designed either for the context in which they are applied or by the models they use, e.g., making use of the unlabeled data available (McCallum and Nigam, 1998; Yu et al, 2006), exploiting the clusters in the data (Dasgupta and Hsu, 2008), diversifying the set of hypotheses (Melville and Mooney, 2004), designing criteria for active learning for collaborative filtering (Jin and Si, 2004; Harpale and Yang, 2008; Boutilier et al, 2003), or adapting the active learning to Gaussian processes (Chu and Ghahramani, 2005a; Brochu et al, 2007).

In this paper we propose an alternative to the standard active learning criteria which makes use of the preference observations collected already from a community of subjects. This criterion, which we call the *Committee criterion*, is thus particularly designed for the multi-task setting that we consider in this paper. The idea behind the Committee criterion is related to the *Query-by-Committee* method from active learning which selects those queries that have maximum disagreement amongst an ensemble of hypotheses. The difference in our case is that the group of subjects, for which the preferences were already learned, plays the role of the ensemble of hypotheses instead of an ensemble of models learned on the same task.

1.3 Notation

Boldface notation is used for vectors and matrices and normal fonts for their components. Upperscripts are used to distinguish between different vectors or matrices and lowerscripts to address their components. The notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used for a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The transpose of a matrix \boldsymbol{M} is denoted by \boldsymbol{M}^T . Capital letters are used for constants and small letters for indices, e.g., $i = 1, \dots, I$.

2 Learning Framework

The idea of using the preference observations from other subjects in order to optimize the process of learning the preferences of a new subject can be basically applied in any preference learning context. In this paper, we restrict ourselves to qualitative preference observations which can be modeled using the probabilistic choice models described in this section.

2.1 Probabilistic Choice Models

Since people are very good in making comparisons, in many real-world applications preferences are learned from experiments in which the subject makes a choice for one of the presented alternatives. Let $X = \{x_1, \dots, x_I\}$ be a set of inputs. Let \mathcal{D} be a set of J observed preference comparisons over instances in X corresponding to a subject,

$$\mathcal{D} = \{(a_j, c_j) | 1 \leq j \leq J, c_j \in \{1, \dots, A\}\} \quad (1)$$

with $a_j = (x_{i_1(j)}, \dots, x_{i_A(j)})$ the alternatives presented and c_j the choice made, $i_1, \dots, i_A : \{1, \dots, J\} \rightarrow \{1, \dots, I\}$ index functions such that $i_1(j)$ represents the input presented first in the j th preference comparison and $c_j = c$ means that $x_{i_c(j)}$ is chosen from the A alternatives presented in the j th comparison. For $A = 2$ this setup reduces to pairwise comparisons between two alternatives.

The main idea behind probabilistic choice models is to assume a latent utility function value $U(x)$ associated with each input x which captures the individual preference of a subject for x . In the ideal case the latent function values are consistent with the preference observations. This means that alternative c is preferred over the other alternatives c' in the j th comparison whenever the utility for c exceeds the utilities for the other alternatives c' , i.e., $U(x_{i_c(j)}) > U(x_{i_{c'}(j)})$. In practice, however, subjects are often inconsistent in their responses. A very inconsistent subject will have a high uncertainty associated with the utility function; this uncertainty is directly taken into account in the probabilistic framework. We define this probabilistic framework by making a standard modeling assumption (Bradley and Terry, 1952; Kanninen, 2002; Glickman and Jensen, 2005) that the probability that the c th alternative is chosen by the subject in the j th comparison follows a multinomial logistic model, which is defined as

$$p(c_j = c | a_j, U) = \frac{\exp[U(x_{i_c(j)})]}{\sum_{c'=1}^A \exp[U(x_{i_{c'}(j)})]} . \quad (2)$$

Efficiently learning preferences reduces to learning the unknown utility function U as accurately and with as few comparisons as possible.

2.2 The Utility Function

The utility function U is a real-valued function, $U : X \rightarrow \mathbb{R}$, which associates with every input $x \in X$ a real number $U(x)$. Each input $x \in X$ is characterized by a set of features, $\phi(x) \in \mathbb{R}^D$. One possible choice for the utility function is to write it as a linear combination of the features,

$$U(x) = \sum_{i=1}^D \alpha_i \phi_i(x) , \quad (3)$$

where α is a vector of weights specific to a subject which captures the importance of each feature of x when evaluating the utility U . The preferences are thus encoded in the vector α and learning the utility function reduces to learning α .

In order to make the definition of the utility function more flexible, we can use a semiparametric model in which the utility function is defined as a linear combination

of basis functions. The basis functions are defined by a kernel κ centered on the data points,

$$U(x) = \sum_{i=1}^I \alpha_i \kappa(x, x_i), \quad (4)$$

where the vector α with dimension I —the number of data points—captures the preferences of the subject. A non-linear utility function can be obtained by using, for example, a Gaussian kernel,

$$\kappa_{\text{Gauss}}(x, x') = \exp\left(-\frac{\ell}{2} \sum_{i=1}^D (\phi_i(x) - \phi_i(x'))^2\right), \quad (5)$$

where ℓ is a length-scale parameter. The two definitions of the utility function from Equations (3) and (4) are similar in the sense that they are both a linear in the parameter. Equation (4) is more suited when the number of features is larger than the number of data points, i.e., $d > n$.

In order to learn the utility function, we use a Bayesian framework in which we treat the vector of parameters α as a random variable. We consider a Gaussian prior distribution over α , $p(\alpha) = \mathcal{N}(\mu, \Sigma)$, which is updated based on the observations from the preference comparisons using Bayes' rule,

$$p(\alpha | \mathcal{D}, \mu, \Sigma) \propto p(\alpha) \prod_{j=1}^J p(c_j | a_j, \alpha), \quad (6)$$

with the likelihood terms of the form given in Equation (2). The choice of the prior will be discussed in the next section. The posterior distribution obtained is approximated to a Gaussian. The Gaussian approximation of the posterior is a good approximation because with few data points the posterior is close to the prior which is a Gaussian, and with many data points the posterior approaches again a Gaussian as a consequence of the central limit theorem (Bishop, 2006). To perform the approximation of the posterior a good choice is to use deterministic methods (e.g., Laplace's method (Mackay, 2002), Expectation Propagation (Minka, 2001)) since they are computationally cheaper than the non-deterministic ones (sampling) and because they are known to be accurate for these types of models (Glickman and Jensen, 2005).

Another approach for modeling the utility function is to learn a weight for each value of a feature/attribute (Guo and Sanner, 2010). We can implement this in our framework by making a transformation of the features. This means that a feature with, for example, 10 possible values will be transformed into 10 binary features, one of the 10 binary features having value 1 if the initial feature takes that value and the rest 0.

2.3 Multi-task Preference Learning

One property which distinguishes preference learning from other learning settings is that in most of the cases preference observations are available from multiple subjects. This property combined with the Bayesian framework allows the transfer of information from other subjects to a new subject. Basically, we use the preference data previously seen from some other subjects to learn an informed prior which will be used as the starting prior when learning the preferences of a new subject. For learning

this informed prior, we use Bayesian hierarchical modeling which assumes that the parameters for individual models are drawn from the same hierarchical prior distribution. Let us assume that we already have preference data available from a group of M subjects. We make the common assumption of a Gaussian prior distribution, $p(\boldsymbol{\alpha}^m) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$, $m = 1, \dots, M$ with the same $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$ for the preference models of all subjects. This prior is updated using Bayes' rule based on the observations from each subject, resulting in a posterior distribution for each individual subject. The hierarchical prior is obtained by maximizing the penalized log-likelihood of all data in a so-called type-II maximum likelihood approach. This optimization is performed by applying the EM algorithm (Gelman et al, 2003), which reduces to the iteration (until convergence) of the following two steps:

E-step: Estimate the sufficient statistics (mean $\boldsymbol{\mu}^m$ and covariance matrix $\boldsymbol{\Sigma}^m$) of the posterior distribution corresponding to each subject m , given the current estimates ($\bar{\boldsymbol{\mu}}^{(t)}$ and $\bar{\boldsymbol{\Sigma}}^{(t)}$) of the hierarchical prior.

M-step: Re-estimate the parameters of the hierarchical prior:

$$\begin{aligned}\bar{\boldsymbol{\mu}}^{(t+1)} &= \frac{1}{M} \sum_{m=1}^M \boldsymbol{\mu}^m, \\ \bar{\boldsymbol{\Sigma}}^{(t+1)} &= \frac{1}{M} \sum_{m=1}^M (\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}^{(t+1)})(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}^{(t+1)})^T + \frac{1}{M} \sum_{m=1}^M \boldsymbol{\Sigma}^m.\end{aligned}\tag{7}$$

Once we have learned the hierarchical prior, $p(\boldsymbol{\alpha}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$, we can use it as an informative prior for the preference model of a new subject in Equation (6).

3 Active Preference Learning

Active learning, also known in the statistics literature as sequential experimental design, is suitable for situations in which labeling points is difficult, time-consuming, and expensive. This is also the case in most of the preference learning settings in which the labels are given by people in an explicit way. The idea behind active learning is that by optimal selection of the training points a greater accuracy can be obtained than by random selection. The scenarios in which active learning can be applied belong to one of the following three categories: *i*) generating *de novo* points for labeling; *ii*) stream-based active learning where the learner decides whether to request the label of a given instance or not; *iii*) pool-based active learning where queries are selected from a large pool of unlabeled data. In this paper we consider pool-based active learning.

We assume that the current model \mathcal{M} of the preference data observed from a subject is a Gaussian distribution $\mathcal{M} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $p(c|a, \mathcal{M})$ be the probability of an observation given the data seen so far and let the new model obtained after incorporating an observation (a, c) be $\mathcal{M}_{(a,c)} \approx \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)})$. There are several strategies for active learning, all being concerned with evaluating the *informativeness* of the unlabeled points. In the following we briefly review these strategies and the way they can be implemented in the learning framework considered here.

1. **Uncertainty Sampling** (Lewis and Gale, 1994). In this strategy an active learner chooses for labeling the example for which the model's predictions are most uncertain. The uncertainty of the predictions can be measured, for example, using

Shannon entropy

$$\text{Uncertainty}(a) = - \sum_c p(c|a, \mathcal{M}) \log p(c|a, \mathcal{M}) . \quad (8)$$

For a binary classifier this strategy reduces to querying points whose prediction probabilities are close to 0.5. Intuitively this strategy aims at finding as fast as possible the decision boundary since this is indicated by the regions where the model is most uncertain.

2. **Variance Reduction** (MacKay, 1992). This strategy, also known in experimental design as D-optimality (Fedorov, 1972; Chaloner and Verdinelli, 1995; Berger, 1994; Ford and Silvey, 1980), chooses as the most informative experiments the ones that give the most reduction in the model's uncertainty. The motivation behind this strategy is a result of (Geman et al, 1992) which shows that the generalization error can be decomposed into three components: *i*) noise (which is independent of the model or training data); *ii*) bias (due to the model); *iii*) model's uncertainty. Since the model cannot influence the noise and the bias components, the future generalization error can only be influenced via the model's variance. Formally, this criterion can be written as

$$\text{Variance}(a) = \sum_c p(c|a, \mathcal{M}) \text{variance}[\mathcal{M}_{(a,c)}] - \text{variance}[\mathcal{M}] . \quad (9)$$

In the setting considered in this paper the the variance of the model is expressed in the covariance of Gaussian distribution. In order to use Equation (9) we need to choose a measure for the variance. We can consider, for example, the log-determinant of the covariance matrix

$$\text{Variance-logdet}(a) = \sum_c p(c|a, \mathcal{M}) \log \det(\boldsymbol{\Sigma}_{(a,c)}) - \log \det(\boldsymbol{\Sigma}) , \quad (10)$$

which is actually minimizing the entropy of the Gaussian random variable representing the current model, or the trace of the covariance matrix

$$\text{Variance-trace}(a) = \sum_c p(c|a, \mathcal{M}) \text{Tr}(\boldsymbol{\Sigma}_{(a,c)}) - \text{Tr}(\boldsymbol{\Sigma}) . \quad (11)$$

3. **Expected Model Change** (Cohn et al, 1996). This strategy chooses as the most informative query the one which when added to the training set would yield the greatest model change. Quantifying the model change depends on the learning framework. For gradient-based optimization the change can be measured via the training gradient, i.e., the vector used to re-estimate parameter values (Settles and Craven, 2008). In the Bayesian framework, the model change can be quantified via a distance measure between the current distribution and the posterior distribution obtained after incorporating the candidate point

$$\text{Change}(a) = \sum_c p(c|a, \mathcal{M}) \text{distance} [\mathcal{M}, \mathcal{M}_{(a,c)}] .$$

A suitable distance for our setting is the Kullback-Leibler divergence between distributions, which for two Gaussians has a closed form solution and can be written

as follows

$$\begin{aligned}
\text{Change-KL}(a) &= \sum_c p(c|a, \mathcal{M}) \text{KL} \left[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)}) \right] \\
&= \sum_c p(c|a, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left[\log \left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}} \right) + \text{Tr} \left(\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma} \right) + \right. \\
&\quad \left. + \left(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}_{(a,c)}^{-1} (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) - n \right]. \tag{12}
\end{aligned}$$

The KL divergence between Gaussians is used by Seeger (2008) to design an efficient sequential experimental design in a setting similar to the one used in this paper.

4. **Query-by-Committee (QBC)** (Seung et al, 1992) is an effective active learning approach that has successfully been applied to many problems. In each iteration QBC *i*) constructs a committee of models based on the current training set, and *ii*) ranks the unlabeled examples according to some measure of disagreement among the committee members. The input with the highest disagreement score is then selected for labeling and added to the training data.

Uncertainty sampling, and QBC and its variants are attractive due to their applicability in various machine learning settings. Variance reduction and expected model change are robust and in many situations they have proved to be the best one can do (Schein and Ungar, 2007). Although more robust, the variance reduction and expected model change strategies are computationally more demanding since for each candidate comparison query and each possible label the posterior distribution induced has to be computed. The posterior distribution cannot be computed analytically and approximations are needed; these approximations are usually costly. In the following we discuss several variants of QBC designed for the setting in which there is available data from multiple subjects (Section 3.1). Furthermore, we show how these variants of QBC can be naturally linked to the hierarchical Bayesian modeling for reducing the computations (Section 3.2). Finally, we show connections between the active learning criteria mentioned above by reducing them to a similar form (Section 3.3).

3.1 QBC for Preference Learning

In this section we will discuss how to adapt QBC to our preference learning setting.

3.1.1 The Committee Members

For the QBC approach to be effective it is important that the committee be made up of consistent and representative models. The main idea in this paper is to exploit the preference learning setting with multiple subjects and use the learned models of other subjects $\mathcal{M}_1, \dots, \mathcal{M}_M$ as the committee members when learning the preferences of a new subject.

After choosing the committee we still have to decide upon a suitable criterion for selecting the next examples. Some measures of disagreement among the committee members appear to be most obvious, and in the following we will consider several alternatives.

3.1.2 Vote Criterion

A simple and straightforward way is to consider the labels assigned by the other subjects, e.g., through the Vote criterion defined as

$$\text{Vote}(a) = \max_c \sum_{m=1}^M \delta(a, c; m), \quad (13)$$

where $\delta(a, c; m) = 1$ if $(a, c) \in \mathcal{D}_m$, and $\delta(a, c; m) = 0$ otherwise. The score $\text{Vote}(a)$ is minimal when the labels assigned by the committee members are equally distributed (total disagreement) and maximal when all members fully agree. There are two problems with this criterion. First, a comparison a may not be labeled by a subject m . This can be overcome if we consider the prediction for a computed based on the learned model of subject m and allow each committee member to ‘vote’ for its winning class. This same idea is implemented in the so-called vote entropy method (Dagan and Engelson, 1995). Second, in practical applications just scoring votes turns out to be sub-optimal. The reason, as also suggested in (McCallum and Nigam, 1998), is that the Vote criterion does not take into account the confidences of the committee members’ predictions.

3.1.3 Committee Criterion

We will use the following notation for the predictive probability corresponding to a subject $m = 1, \dots, M$

$$p_m(c|a) \equiv p(c|a, \mathcal{M}_m).$$

The predictive probability can be computed either by taking into account the entire distribution $\mathcal{M}_m = \mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$

$$p_m(c|a) = \int p(c|a, \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{\alpha} | \boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m) d\boldsymbol{\alpha},$$

or, for computational reasons, we can consider only a point estimate for \mathcal{M}_m , for example, the mean of the Gaussian distribution, and use it in Equation (2)

$$p_m(c|a) = p(c|a, \boldsymbol{\mu}^m). \quad (14)$$

Inspired by (McCallum and Nigam, 1998), we propose to measure disagreement by taking the average prediction of the entire committee and computing the average Kullback-Leibler (KL) divergence of the individual predictions from the average:

$$\text{Committee}(a) = \sum_{m=1}^M \frac{1}{M} \text{KL}[\bar{p}(\cdot|a) || p_m(\cdot|a)], \quad (15)$$

with $\bar{p}(\cdot|x)$ the average predictive probability of the entire committee, which will be more precisely defined in Section 3.2.

The KL divergence for discrete probabilities is defined as

$$\text{KL}[p_1(\cdot|a) || p_2(\cdot|a)] = \sum_c p_1(c|a) \log \left(\frac{p_1(c|a)}{p_2(c|a)} \right).$$

The KL divergence can be seen as a distance between probabilities, where we abused the notion of distance, since the KL-divergence is not symmetric, $\text{KL}[p_1||p_2] \neq \text{KL}[p_2||p_1]$. This drawback of the KL-divergence can be overcome by considering a symmetric measure, for example, $\text{KL}[p_1||p_2] + \text{KL}[p_2||p_1]$. In (McCallum and Nigam, 1998), the disagreement is computed between committee members constructed based on the current model, i.e., the committee changes with every update and the criterion has to be recomputed with every update. A committee of models learned on different tasks is fixed and thus selecting examples solely based on it leads to a fixed instead of an active design: all examples can be ranked beforehand (the same applies to the Vote criterion defined above).

To arrive at an active design and take into account the current model, we propose a small modification, based on the following intuition. Querying examples on which the committee members disagree makes sense, because it will force the current model to make a choice between options that, according to the committee members, are reasonably plausible. However, when the current model on a particular example already “made up its mind”, i.e., deviates substantially from the average prediction of the committee based on what it learned from other input/output pairs, it makes no sense to still query that example, even though the committee members might disagree. Taking into account this consideration, we propose the Committee criterion which assigns a score to a candidate query comparison a through

$$\text{Committee}(a) = \frac{1}{M} \sum_{m=1}^M \text{KL}[\bar{p}(\cdot|a)||p_m(\cdot|a)] - \gamma \text{KL}[\bar{p}(\cdot|a)||p(\cdot|a)], \quad (16)$$

with $p(\cdot|a)$ the current model’s predictive probability based on the data seen so far and γ a parameter that accounts for the degree of similarity between subjects. According to this Committee criterion, the most interesting experiments are those on which the other models disagree (the first term on the righthand side of Equation (16)), with the current model (still) undecided (the second term on the righthand side of Equation (16)).

An advantage of the Committee criterion is its computational efficiency: the first term on the righthand side of Equation (16) as well as the average predictive probability can be computed beforehand. The Committee criterion does require computation of the predictive probabilities corresponding to the current model, but this is the least one could expect from an active design. This is to be compared with the QBC criterion, which requires constructing new committee members with each update, and D-optimal experimental design, which calls for keeping track of variances.

Note that we have not made any restriction so far with respect to the probabilistic models used in the active learning design. In the following we will consider only the log-linear models introduced in Section 2. They have some nice properties, which simplify the computation of the Committee criterion (Section 3.2), and provide a natural link to hierarchical Bayesian modeling (Section 2.3). The general idea, of using the already learned models from the other tasks as the committee members in a QBC-like approach, is of course also applicable to other models.

3.2 Average Probability

For linear utility functions the likelihood function defined in Equation (2) is a log-linear model (Christensen, 1997). The log-odds of the model are linear in the parameter.

Let $p_m(c|a)$ be the predictive probability defined in Equation (14). We define the average predictive probability of the committee, $\bar{p}(c|a)$, as the prediction probability that is closest to the prediction probabilities of the members:

$$\bar{p}(c|a) \equiv \operatorname{argmin}_{p(c|a)} \sum_{m=1}^M \frac{1}{M} \text{KL}[p(c|a) || p_m(c|a)] . \quad (17)$$

The solution is the so-called logarithmic opinion pool (Bordley, 1982)

$$\bar{p}(c|a) = \frac{1}{Z(a)} \prod_{m=1}^M [p_m(c|a)]^{\frac{1}{M}} = \frac{1}{Z(a)} \exp \left(\frac{1}{M} \sum_{m=1}^M \log p_m(c|a) \right) , \quad (18)$$

with $Z(a)$ a normalization constant

$$Z(a) = \sum_c \prod_{m=1}^M [p_m(c|a)]^{\frac{1}{M}} .$$

For log-linear models, the logarithmic opinion pool boils down to a simple averaging of model parameters:

$$\bar{p}(c|a) = p(c|a, \bar{\mu}) \quad \text{with} \quad \bar{\mu} = \frac{1}{M} \sum_{m=1}^M \mu^m . \quad (19)$$

This natural combination between log-linear models and logarithmic opinion pools is the advantage of using the logarithmic opinion pool instead of the linear opinion pool used in (McCallum and Nigam, 1998).

As can be seen from the EM updates in Equation (7), the average $\bar{\mu}$ in the logarithmic opinion pool is then precisely the mean of the learned hierarchical prior. Summarizing, once we have learned a hierarchical prior from the data available for subjects 1 through M using the EM algorithm, we can start off the new model $M + 1$ from this prior (as is normally done in hierarchical Bayesian learning). On top of this, the same EM algorithm gives us the information we need to compute the Committee criterion that can be used subsequently to select new inputs to label.

3.3 Similarities between Criteria

In this section we investigate how similar the active learning criteria are and how can they be related. We analyze the modifications induced to the model by the criteria after updating the probability model to incorporate the information from new training points. A single update induces a small change in the posterior distribution, and this allows for Taylor expansions, keeping only the lowest non-zero contribution. In the following we present the main results of the approximations while the details can be found in the appendix.

Assuming that the updates of the posterior distribution for each alternative a and choice c lead to small changes in the model \mathcal{M} we can approximate the criteria to the form

$$\sum_c p(c|a, \alpha) g(c|a, \alpha)^T Q g(c|a, \alpha) , \quad (20)$$

for some vector α and matrix Q and with g the gradient of the log-probabilities

$$g(c|a, \alpha) \equiv \frac{\partial \log p(c|a, \alpha)}{\partial \alpha}.$$

Specifically, it can be shown using Taylor expansion that the Variance-logdet can be approximated as, (see Lemma 2 from appendix for the derivation)

$$\text{Variance-logdet}(a) \approx \sum_c p(c|a, \mu) g(c|a, \mu)^T \Sigma g(c|a, \mu). \quad (21)$$

where μ and Σ represent the mean and covariance of the Gaussian posterior distribution. Using the same type of approximation, the Committee criterion can be approximated to the same form, (see Lemma 3 from appendix for the derivation):

$$\text{Committee}(a) \approx \frac{1}{2} \sum_c p(c|a, \bar{\mu}) g(c|a, \bar{\mu})^T \tilde{\Sigma} g(c|a, \bar{\mu}), \quad (22)$$

with $\bar{\mu}$ the hierarchical prior mean and

$$\tilde{\Sigma} \equiv \frac{1}{M} \sum_{m=1}^M (\mu^m - \bar{\mu})(\mu^m - \bar{\mu})^T - (\mu - \bar{\mu})(\mu - \bar{\mu})^T.$$

Furthermore, it can also be proved that Variance-trace and Change-KL can be approximated to the same form from (20), namely

$$\text{Variance-trace}(a) \approx \sum_c p(c|a, \mu) g(c|a, \mu)^T \Sigma^2 g(c|a, \mu), \quad (23)$$

$$\frac{1}{2} \text{Change-KL}(a) \approx \text{Variance-logdet}(a). \quad (24)$$

We will focus on the differences between the Variance-logdet criterion (considered as the reference) and the Committee criterion. The differences between their approximations are as follows.

1. The gradients $g(c|a, \cdot)$ are evaluated at different points: the prior hierarchical mean $\bar{\mu}$ and the current posterior mean μ . This effect is small since μ is still close enough to $\bar{\mu}$ for a sufficiently accurate approximation of the gradients, in particular at the start of the learning when selecting the right points to label is more important.
2. The current posterior variance Σ is replaced by $\tilde{\Sigma}$. The effect of the precise weighting of the gradients is not so important, and again, at the beginning of learning $\tilde{\Sigma}$ is close to Σ .

The way in which experiments are selected is more important at the beginning of the learning process, when μ is still close to the prior mean $\bar{\mu}$, and $\tilde{\Sigma}$ to Σ .

4 Experimental Evaluation

4.1 Data Sets

We used two data sets related to the preferences of people.

The first data set is related to people’s preferences for sound quality. It consists of evaluations of sound quality from 32 people. Each person was subjected to 576 paired-comparison listening experiments of the form (a, c) , where the alternatives are $a = (x_1, x_2)$ and x_1 and x_2 represent one sound sample processed with two different settings of the hearing-aid parameters, and the choice $c = \{1, 2\}$ denotes which of the two alternatives was preferred by the user. More details about this data can be found in (Arehart et al, 2007).

The second data set is related to people’s preferences for art images. The preferences were collected from a web-based survey in which 190 subjects participated and in which 642 images were available for rating. A number of images were presented to each subject, and the subject was asked for her preference with respect to each of them: like/dislike. On average, each subject rated around 90 images; we considered the 32 subjects who rated more than 120 images. Each image is described by a 275-dimensional feature vector, such as, color, shape and texture. For computational efficiency reasons we used a subset of the 10 most informative features which was measured by averaging the correlations between the feature and the observations corresponding to a subject. More details about this data can be found in (Yu et al, 2003).

4.2 Performance

Our experiments use a leave-one-out scheme in which each subject was considered once as the current/new subject for which the preferences need to be learned in order to make a recommendation. The preference data from the rest of the subjects was used first, to learn an informative prior for the current subject and second, for selecting the preference data to be presented to the current subject for labeling. The goal is to minimize the size of the training data needed to accurately learn the preferences of the current subject. This is one of the bottle-necks in most of the preference learning situations since obtaining labeled training data is expensive.

In order to quantify the improvement obtained in learning preferences using the optimizations proposed in this paper we compared the proposed Committee criterion and the reference Variance-logdet criterion with random selection. For each subject the learning started with the hierarchical prior learned from the data of the other subjects. The hierarchical prior was learned from 20 randomly selected data points per subject for the audio data set and from 15 randomly selected data points per subject for the art data set. The data for the current subject was split into 5 folds, 1 fold was used for learning and the rest was used for testing. In the audio data, for each subject 5 folds were made like in a cross-validation setting. In the art data each subject rated at least 120 images and the folds were made by randomizing the data associated with one subject 5 times and selecting 100 data points in the training and the rest for testing. The training data was used as a pool out of which points were selected for labeling either randomly or optimally using one of the active learning criteria. The hierarchical prior was updated based on these data points. After every update predictions were made on the test set using the current model; the accuracy of the predictions on the

test data was used as a measure of how much we learned about the subject preferences. The results were averaged over the 5 splits and over the subjects.

Figure 1 shows the performance of the Committee criterion on the left and Variance-logdet criterion on the right vs random selection on the audio data. The plots show the prediction accuracy (on the y -axis) as a function of the number of updates from the hierarchical prior (on the x -axis). The shaded region indicates the accuracy of 10 random selection runs. Figure 2 shows the same comparison but on the art data set. For both data sets the informative prior improves the predictions at the beginning when no preference observations have been observed for the new subject. The hierarchical prior already gives an accuracy of almost 0.7 for the audio data and of 0.6 for the art data at the beginning of learning.

The learning process can be further optimized by optimally selecting the points to be labeled by the new subject. For the audio data we used the Committee criterion with $\gamma = 0$ since the subjects in the committee are quite similar between each other, which is also suggested by the small error bars. For the art data which has a higher variability between subjects the Committee criterion with $\gamma = 1$ performs slightly better than the Committee criterion with $\gamma = 0$. The audio data contains a few very informative data points and some which are not informative. In some cases the difference between the 2 sound samples presented in an experiment is so small that the subject cannot hear any difference. Such experiments are not informative because a subject’s answer is almost random and does not bring any information with respect to the subject’s preferences. The active learning criteria avoid selecting these type of experiments and obtain better performance than random selection. The preferences of people for art images are more difficult to predict, since preferences do not depend on some low-level characteristics of the image, like texture, color, etc. This is why the accuracy obtained on the art data is less than the one obtained on the audio data. The Committee criterion shows very similar performance to the Variance-logdet criterion on the audio data. The Variance-logdet criterion does slightly better on the art data. On both data sets, a higher accuracy can be obtained by optimally selecting points to label than by just random selection.

4.3 Similarities between Criteria

In order to test empirically the approximations and similarities from Section 3.3 we computed the Spearman rank correlations between the scores assigned by the criteria when evaluating the informativeness of the data points. For each subject the learning started with the hierarchical prior learned from the data of the other subjects. This prior was updated by taking into account the information from 20 randomly selected data points for both data sets. After these updates, we computed the scores assigned by each of the active learning criteria to 50 randomly chosen data points. We computed the Spearman rank correlations between these scores for each pair of criteria. Figure 3 shows the Spearman rank correlation coefficients for each pair of criteria; the darker the color the closer to 1 the correlations are and the stronger the similarity between the two criteria. There are several observations to be made from this figure: *i*) One can notice a darker square on the left-down part of the figures, both for the audio and art data set. This square involves the Variance-logdet, Change-KL, Variance-trace, and Committee criteria. The correlations between each pair of them are close to 1 which suggests that these criteria perform in practice very similar. This is also what the

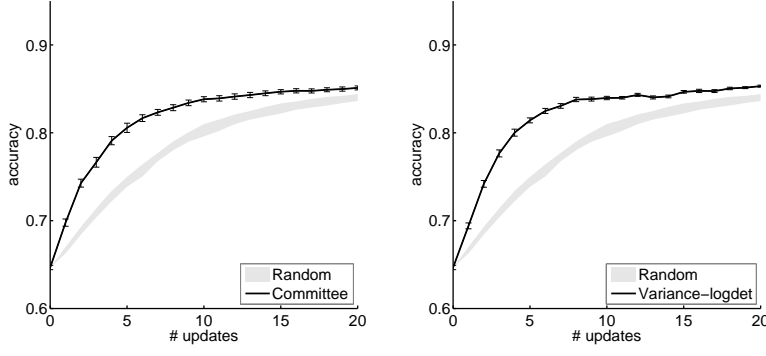


Fig. 1: Performance of the Committee criterion on the left and Variance-logdet criterion on the right vs random selection for the audio data set. The plots show the prediction accuracy (on the y -axis) as a function of the number of updates from the hierarchical prior (on the x -axis). The error bars give the standard deviation of the mean accuracy, averaged over the 32 subjects. The shaded region shows the range of 10 random strategies.

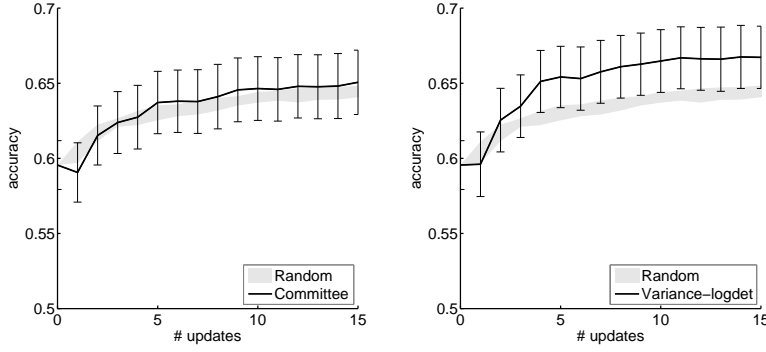


Fig. 2: Performance of the Committee criterion on the left and Variance-logdet criterion on the right vs random selection for the art data set. The plots show the prediction accuracy (on the y -axis) as a function of the number of updates from the hierarchical prior (on the x -axis). The error bars give the standard deviation of the mean accuracy, averaged over the 32 subjects. The shaded region shows the range of 10 random strategies.

theory from Section 3.3 suggests by approximating these criteria to a similar form. *ii)* The Variance-logdet and Change-KL criterion have the Spearman rank correlation very close to 1. Their approximations are proven to be equivalent in Lemma 5 in the Appendix. These two observations also suggest that the approximations of the Variance-logdet and Change-KL are very accurate. *iii)* The Vote criterion performs in some situations randomly since when the number of subjects is much smaller than the number of data points considered, the scores assigned by the Vote criterion are the same to most of the experiments. *iv)* The Uncertainty criterion is most different from the others.

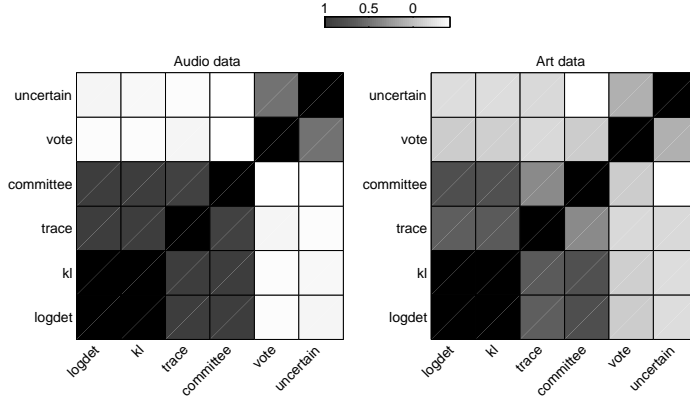


Fig. 3: Spearman rank correlation coefficients for the scores assigned by the active learning criteria. On the left the correlations computed for audio data and on the right for art data. The darker the color, the higher the correlations, suggesting that the criteria are very similar. The lighter the color, the lower the correlations, suggesting that the criteria are not similar.

5 Conclusions and Discussions

This study started with the premise that when preference learning is involved preference data is available from multiple subjects. Preferences are typically learned by asking questions or doing experiments with a subject. A major problem in this approach is that obtaining new observations is often a costly and/or time consuming process, but the learning process can be optimized by using the responses from other subjects.

Our approach combines hierarchical Bayesian modelling (to learn from related tasks) with active learning (to learn efficiently by selecting informative examples). A hierarchical prior was used from which model parameters were sampled in order to enforce a similar structure for the utility functions of each individual subject. The Committee criterion we proposed for active learning was shown to behave similarly to the well established criteria from experimental design, but is computationally more efficient. The idea behind the Committee criterion is to have the data available from other subjects collaborate in order to select the most informative experiments to perform with a new subjects. We showed that the conceptual idea behind Query-by-Committee also works with a committee of people.

We have illustrated our approach for optimizing the learning of preferences using real-world data as examples. We used an audio data with various distortions of the sounds, where the subjects had a larger degree of agreement and we used an art data where the subjects had a lesser degree of agreement as art is more a matter of personal taste.

There are several directions for extending our approach. *i)* A natural extension would be to relax the constraint that the data from the other subjects is given. Thus we could learn the models of all subjects in parallel, i.e., making the Disagreement active. Consider the situation in which the preferences of a group of subjects are learned in the same time. We obtain something similar to interleaving EM with active learning: in the E-step the data from other subjects is used to fill-in the hidden labels and select the most informative data point to label for the current subject, and in the M-step

the data of the current subject is used for learning a more accurate hierarchical model. In the same direction, another extension would be to weigh the contributions of the committee members. The intuition is that those members with smaller uncertainty in their models get larger weights. McCallum and Nigam (1998) used the idea of interleaving EM with active learning to increase the performance by using the unlabeled data. *ii*) Another direction worth investigating is a non-myopic design, similar to the one proposed by Boutilier (2002). A non-myopic design looks more than just one step in advance when evaluating the informativeness of a data point. It is theoretically closer to the best possible design but computationally much more expensive. Due to the computational complexity involving a non-myopic design, we discussed all the active learning criteria from a myopic perspective, however, a non-myopic perspective can be applied to all of them. *iii*) The end goal of learning the preference of a person is to make recommendations about an item he/she would like. In this paper we focused on accurately learning the utility function. The criteria discussed in this paper could be adapted to the setting in which we focus on finding the item which maximizes the utility function, similar to the criteria for finding the maximum of the utility function proposed in (Groot et al, 2010). *iv*) In this paper we used log-linear models and Gaussian distributions to model the preference data. The same idea, of using models learned on data from different subjects (or scenarios) to actively select examples for a new subject, can be applied to other models and starting from different priors as well, although the mathematics will be a bit more involved and less intuitive. In particular, considering a mixture of Gaussians as the prior may still be feasible and may lead to an active learning strategy that tries to find those examples that can best discriminate to which mixture component the current model belongs.

Appendix

In this appendix we prove the equivalences between the active learning criteria stated in Section 3.3. We show that these criteria can be approximated to the same form, namely

$$\sum_c p(c|a, \boldsymbol{\alpha}) \mathbf{g}(c|a, \boldsymbol{\alpha})^T \mathbf{Q} \mathbf{g}(c|a, \boldsymbol{\alpha}),$$

for some vector $\boldsymbol{\alpha}$ and matrix \mathbf{Q} . The difference between the approximations for different criteria is the point $\boldsymbol{\alpha}$ in which the gradients and the probabilities are evaluated and the weighting matrix of the gradients \mathbf{Q} .

We consider probabilistic choice models of the form given in Equation (2), which by using the definition of the utility function from Equations (3) and (4) can be rewritten as

$$p(c|a, \boldsymbol{\alpha}) = \frac{\exp \left[\sum_{i=1}^D \phi_i(x_c) \alpha_i \right]}{Z(\boldsymbol{\alpha}, a)} \quad \text{with} \quad Z(\boldsymbol{\alpha}, a) \equiv \sum_{c'} \exp \left[\sum_i \phi_i(x_{c'}) \alpha_i \right].$$

We define the derivatives of the log probabilities

$$\mathbf{g}(c|a, \boldsymbol{\alpha}) \equiv \frac{\partial \log p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}, \quad \mathbf{H}(c|a, \boldsymbol{\alpha}) \equiv \frac{\partial^2 \log p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}.$$

We first prove a lemma which states a relationship between the Hessian and the Fisher matrices which will be used in further proofs.

Lemma 1 *For any input a and vector $\boldsymbol{\alpha}$ we have the following relationship between the Hessian and the Fisher matrices:*

$$\sum_c p(c|a, \boldsymbol{\alpha}) \mathbf{H}(c|a, \boldsymbol{\alpha}) = - \sum_c p(c|a, \boldsymbol{\alpha}) \mathbf{g}(c|a, \boldsymbol{\alpha}) \mathbf{g}(c|a, \boldsymbol{\alpha})^T,$$

Proof. We use shorthand notation $p_c = p(c|a, \boldsymbol{\alpha})$, $g_{cj} = g_j(c|a, \boldsymbol{\alpha})$, $\phi_{ci} = \phi_i(x_c)$, omitting the dependencies on a and $\boldsymbol{\alpha}$.

From $\log p_c = \sum_j \phi_{cj} \alpha_j - \log Z$, it is easy to see that

$$g_{cj} = \phi_{cj} - \frac{\partial \log Z}{\partial \alpha_j}, \quad H_{c,ij} = -\frac{\partial^2 \log Z}{\partial \alpha_i \partial \alpha_j}.$$

Furthermore,

$$\begin{aligned} \frac{\partial \log Z}{\partial \alpha_j} &= \frac{1}{Z} \frac{\partial Z}{\partial \alpha_j} = \frac{1}{Z} \sum_c \exp \left[\sum_{j'} \phi_{cj'} \alpha_{j'} \right] \phi_{cj} = \sum_c p_c \phi_{cj}, \\ \frac{\partial^2 \log Z}{\partial \alpha_i \partial \alpha_j} &= \sum_c \phi_{cj} \left[\frac{\exp(\sum_{j'} \phi_{cj'} \alpha_{j'}) \phi_{ci} Z - \frac{\partial Z}{\partial \alpha_i} \exp(\sum_{j'} \phi_{cj'} \alpha_{j'})}{Z^2} \right] \\ &= \sum_c \phi_{cj} [\phi_{ci} p_c - \sum_{c'} (p_{c'} \phi_{c'i}) p_c] = \sum_c p_c \phi_{cj} \phi_{ci} - \sum_c p_c \phi_{cj} \sum_{c'} p_{c'} \phi_{c'i}, \end{aligned}$$

and thus

$$\begin{aligned} g_{cj} &= \phi_{cj} - \sum_{c'} p_{c'} \phi_{c'j}, \\ H_{c,ij} &= -\sum_{c'} p_{c'} \phi_{c'j} \phi_{c'i} + \sum_{c'} p_{c'} \phi_{c'j} \sum_{c''} p_{c''} \phi_{c''i} = H_{ij}. \end{aligned} \quad (25)$$

Note that the second derivative is in fact independent of c . We then have

$$\begin{aligned} \sum_c p_c H_{c,ij} &= \sum_c p_c H_{ij} = H_{ij} = -\sum_c p_c \phi_{ci} \phi_{cj} + \sum_c p_c \phi_{ci} \sum_{c'} p_{c'} \phi_{c'j} \\ &= -\sum_c p_c \left(\phi_{ci} - \sum_{c'} p_{c'} \phi_{c'i} \right) \left(\phi_{cj} - \sum_{c'} p_{c'} \phi_{c'j} \right) = -\sum_c p_c g_{ci} g_{cj}. \end{aligned}$$

□

The following lemma proves the approximation of the Variance-logdet criterion from Equation (21).

Lemma 2 *In a first order approximation, assuming that $\boldsymbol{\Sigma}_{(a,c)}$ is close to $\boldsymbol{\Sigma}$, we can simplify*

$$\text{Variance-logdet}(a) \approx \sum_c p(c|a, \boldsymbol{\mu}) \mathbf{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma} \mathbf{g}(c|a, \boldsymbol{\mu}).$$

Proof.

In a first order approximation we have

$$\boldsymbol{\Sigma}_{(a,c)}^{-1} \approx \boldsymbol{\Sigma}^{-1} - \frac{\partial^2 \log p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \quad (26)$$

where we ignored the change from the old $\boldsymbol{\alpha}$ to a new MAP solution depending on c and a .

For a matrix A and ϵ small compared to A , the following holds (see, for example, Boyd and Vandenberghe (2004), pag. 642)

$$\log \det(A + \epsilon \mathbf{I}) \approx \log \det(A) + \text{Tr}[A^{-1} \epsilon]. \quad (27)$$

Assuming $\boldsymbol{\Sigma}_{(a,c)}^{-1}$ is close to $\boldsymbol{\Sigma}^{-1}$ we can use Equation (26) in Equation (27) to obtain

$$\log \det \boldsymbol{\Sigma}_{(a,c)}^{-1} \approx \log \det \boldsymbol{\Sigma}^{-1} - \text{Tr}[\boldsymbol{\Sigma} \mathbf{H}(c|a, \boldsymbol{\mu})]. \quad (28)$$

The probability that the subject gives the response c when presented the alternatives a follows by integrating $p(c|a, \boldsymbol{\alpha})$ over the current posterior. We make a second order Taylor expansion of $p(c|a, \boldsymbol{\alpha})$ around the point $\boldsymbol{\mu}$:

$$\begin{aligned} p(c|a) &= \int d\boldsymbol{\alpha} p(c|a, \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\approx \int d\boldsymbol{\alpha} \left[p(c|a, \boldsymbol{\alpha}) + (\boldsymbol{\alpha} - \boldsymbol{\mu})^T \frac{\partial p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} + \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\mu})^T \frac{\partial^2 p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} (\boldsymbol{\alpha} - \boldsymbol{\mu}) \right] \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= p(c|a, \boldsymbol{\mu}) + \frac{1}{2} \text{Tr} \left[\boldsymbol{\Sigma} \frac{\partial^2 p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \right]. \end{aligned}$$

The first order term cancels since the gradient is zero at the maximum solution $\boldsymbol{\alpha} = \boldsymbol{\mu}$. In a lowest order approximation we can ignore the correction upon $p(c|a, \boldsymbol{\alpha})$ to obtain

$$\begin{aligned} \text{Variance-logdet}(a) &= - \sum_c p(c|a, \boldsymbol{\mu}) \log \det \boldsymbol{\Sigma}_{(a,c)} + \log \det \boldsymbol{\Sigma} \\ &= - \sum_c p(c|a, \boldsymbol{\mu}) [-\log \det(\boldsymbol{\Sigma}_{(a,c)}^{-1}) + \log \det(\boldsymbol{\Sigma}^{-1})] \\ &\approx - \sum_c p(c|a, \boldsymbol{\mu}) \text{Tr}[\boldsymbol{\Sigma} \mathbf{H}(c|a, \boldsymbol{\mu})], \end{aligned}$$

where for the last approximation we used the approximation from Equation (28). Lemma 1 then gives the result. \square

The following lemma proves the approximation of the Committee criterion from Equation (22).

Lemma 3 *In a lowest order approximation the Committee criterion can be written as*

$$\frac{1}{2} \text{Committee}(a) = \frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \mathbf{g}(c|a, \bar{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}} \mathbf{g}(c|a, \bar{\boldsymbol{\mu}}),$$

where $\bar{\boldsymbol{\mu}}$ is the mean of the hierarchical prior learned from the other subjects and

$$\tilde{\boldsymbol{\Sigma}} \equiv \frac{1}{M} \sum_{m=1}^M (\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}})^T - (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T.$$

We make a second order Taylor expansion of the KL divergences from the definition of the Committee criterion in Equation (16)

$$\text{KL}[\bar{p}(\cdot|a)||p(\cdot|a)] = \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \log \left[\frac{p(c|a, \bar{\boldsymbol{\mu}})}{p(c|a, \boldsymbol{\mu})} \right],$$

around the point $\bar{\boldsymbol{\mu}}$.

The first order term is:

$$- \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \frac{\partial \log p(c|a, \bar{\boldsymbol{\mu}})}{\partial \bar{\boldsymbol{\mu}}} \Big|_{\bar{\boldsymbol{\mu}}=\bar{\boldsymbol{\mu}}} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T = \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \mathbf{g}(c|a, \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T$$

which cancels since from Equation (25),

$$\begin{aligned} g_{cj} &= \phi_{cj} - \sum_{c'} p_{c'} \phi_{c'j} \\ \Rightarrow \sum_c p_c g_{cj} &= \sum_c p_c \phi_{cj} - \sum_c p_c \sum_{c'} p_{c'} \phi_{c'j} = \sum_c p_c \phi_{cj} - \sum_{c'} p_{c'} \phi_{c'j} = 0, \end{aligned}$$

thus $\sum_c p(c|a, \boldsymbol{\alpha}) \mathbf{g}(c|a, \boldsymbol{\alpha}) = 0$.

The second order term can be rewritten by using the result from Lemma 1 as:

$$\begin{aligned} -\frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \mathbf{H}(c|a, \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \\ = \frac{1}{2} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \mathbf{g}(c|a, \bar{\boldsymbol{\mu}}) \mathbf{g}(c|a, \bar{\boldsymbol{\mu}})^T (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \\ = \frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \mathbf{g}(c|a, \bar{\boldsymbol{\mu}})^T (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \mathbf{g}(c|a, \bar{\boldsymbol{\mu}}). \end{aligned}$$

Since the other terms cancels, we obtain that the KL-divergence between the predictive probabilities can be approximated as

$$\text{KL}[\bar{p}(\cdot|a) || p(\cdot|a)] = \frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \mathbf{g}(c|a, \bar{\boldsymbol{\mu}})^T (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \mathbf{g}(c|a, \bar{\boldsymbol{\mu}}).$$

Making this approximation for all the KL-divergences in the definition of the Committee criterion from Equation (16) and computing the sum we obtain the result stated in the lemma. \square

The following lemma proves the approximation of the Variance-trace criterion from Equation (23).

Lemma 4 *In a first order approximation, the Variance-trace criterion boils down to*

$$\text{Variance-trace}(a) = \sum_c p(c|a, \boldsymbol{\mu}) \mathbf{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma}^2 \mathbf{g}(c|a, \boldsymbol{\mu}).$$

Proof. We have

$$\boldsymbol{\Sigma}_{(a,c)} = \left(\boldsymbol{\Sigma}_{(a,c)}^{-1} \right)^{-1} = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_{(a,c)}^{-1} - \boldsymbol{\Sigma}^{-1} \right)^{-1} \approx -\boldsymbol{\Sigma} \left[\boldsymbol{\Sigma}_{(a,c)}^{-1} - \boldsymbol{\Sigma}^{-1} \right] \boldsymbol{\Sigma}.$$

Use of Equation (26) and Lemma 1 gives the result. \square

The following lemma proves the approximation of the Change-KL criterion from Equation (24).

Lemma 5 *In a first order approximation, assuming that $\boldsymbol{\Sigma}_{(a,c)}$ is close to $\boldsymbol{\Sigma}$, we have*

$$\text{Change-KL}(a) = \text{Variance-logdet}(a),$$

i.e., the two criteria are indistinguishable.

Proof. We evaluate the terms of the Change-KL criterion one by one,

$$\text{Change-KL}(a) = \sum_c p(c|a) \left[\log \left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}} \right) + \text{Tr} \left(\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma} \right) + (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{(a,c)}^{-1} (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) - n \right]$$

The first term gives

$$\sum_c p(c|a) \log \left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}} \right) = -\text{Variance-logdet}(a).$$

The second term

$$\begin{aligned} \sum_c p(c|a) \text{Tr} \left[\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma} \right] &= \sum_c p(c|a) \text{Tr} [(\boldsymbol{\Sigma}^{-1} - \mathbf{H}(c|a, \boldsymbol{\mu})) \boldsymbol{\Sigma}] \\ &= n - \sum_c p(c|a) \text{Tr} [\mathbf{H}(c|a, \boldsymbol{\mu}) \boldsymbol{\Sigma}] = n - \sum_c p(c|a, \boldsymbol{\mu}) \text{Tr} [\boldsymbol{\Sigma} \mathbf{H}(c|a, \boldsymbol{\mu})] \\ &= n + \text{Variance-logdet}(a) \end{aligned}$$

In the same lowest order, we obtain for the third term

$$\begin{aligned} \sum_c p(c|a) (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{(a,c)}^{-1} (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) &\approx \\ &\approx \sum_c p(c|a, \boldsymbol{\mu}) \mathbf{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma} \mathbf{g}(c|a, \boldsymbol{\mu}) = \text{Variance-logdet}(a). \end{aligned}$$

Collection of all the terms then gives the result. \square

References

- Aiolfi F, Sperduti A (2004) Learning preferences for multiclass problems. In: *Advances in Neural Information Processing Systems 17*, MIT Press, pp 17–24
- Arehart K, Kates J, Anderson C, Harvey Jr L (2007) Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 122(2):1150–1164
- Argyriou A, Micchelli C, M P (2008) When is there a representer theorem? vector versus matrix regularizers. *JMLR*
- Bakker B, Heskes T (2003) Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research* 4:83–99
- Berger M (1994) D-optimal sequential sampling designs for item response theory models. *Journal of Educational and Behavioral Statistics* 19:43–56
- Birlutiu A, Groot P, Heskes T (2009) Multi-Task Preference Learning with an Application to Hearing Aid Personalization. In: *Neurocomputing*, DOI 10.1016/j.neucom.2009.11.025
- Bishop C (2006) *Pattern Recognition and Machine Learning*. Springer
- Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022, DOI 0.1162/jmlr.2003.3.4-5.993
- Blythe J (2002) Visual exploration and incremental utility elicitation. In: *Eighteenth national conference on Artificial intelligence*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp 526–532
- Bordley R (1982) A multiplicative formula for aggregating probability assessments. *Management Science* 28(10):1137–1148
- Boutilier C (2002) A pomdp formulation of preference elicitation problems. In: *Eighteenth national conference on Artificial intelligence*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp 239–246
- Boutilier C, Zemel R, Marlin B (2003) Active collaborative filtering. In: *Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp 98–106
- Boyd S, Vandenberghe L (2004) *Convex Optimization*. Cambridge University Press
- Bradley R, Terry M (1952) Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika* 39:324–345
- Brochu E, de Freitas N, Ghosh A (2007) Active preference learning with discrete choice data. In: *Advances in Neural Information Processing Systems*
- Caruana R, Baluja S, Mitchell T (1996) Using the future to sort out the present: Rankprop and multitask learning for medical risk evaluation. In: *Advances in Neural Information Processing Systems 8*, pp 959–965
- Chajewska U, Koller D, Parr R (2000) Making rational decisions using adaptive utility elicitation. In: *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pp 363–369
- Chaloner K, Verdinelli I (1995) Bayesian experimental design: A review. *Statistical Science* 10:273–304
- Christensen R (1997) *Log-Linear Models and Logistic Regression*, 2nd edn. Springer-Verlag
- Chu W, Ghahramani Z (2005a) Extensions of Gaussian processes for ranking: semi-supervised and active learning. In: *NIPS 2005 Workshop on Learning to Rank*, Whistler, BC
- Chu W, Ghahramani Z (2005b) Preference Learning with Gaussian Processes. In: *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany
- Clyde M, Muller P, Parmigiani G (1993) Optimal designs for heart defibrillators
- Cohn D, Ghahramani Z, Jordan M (1996) Active learning with statistical models. *Journal of Artificial Intelligence Research* 4:129–145

- Crammer K, Singer Y (2001) Pranking with ranking. In: *Advances in Neural Information Processing Systems 14*, MIT Press, pp 641–647
- Dagan I, Engelson SP (1995) Committee-based sampling for training probabilistic classifiers. In: *Proceedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann, pp 150–157
- Dasgupta S, Hsu D (2008) Hierarchical sampling for active learning. In: *ICML '08: Proceedings of the 25th international conference on Machine learning*, ACM, New York, NY, USA, pp 208–215, DOI 10.1145/1390156.1390183
- Doyle J (2004) Prospects for preferences. *Computational Intelligence* 20(2):111–136
- Dror H, Steinberg D (2008) Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association* (103):288–298
- Evgeniou T, Micchelli C, Pontil M (2005) Learning multiple tasks with kernel methods. *Journal of Machine Learning Research* 6:615–637
- Fedorov V (1972) *Theory of Optimal Experiments*. Academic Press
- Ford I, Silvey S (1980) A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika* 67:381–388
- Freund Y, Shamir E, Tishby N (1997) Selective sampling using the Query by Committee algorithm. In: *Machine Learning*, pp 133–168
- Furnkranz J, Hullermeier E (2003) Pairwise preference learning and ranking. In: *Proceedings of the 14th European Conference on Machine Learning*, Springer-Verlag, pp 145–156
- Gelman A, Carlin J, Stern H, Rubin D (2003) *Bayesian Data Analysis*, Second Edition. Chapman & Hall/CRC
- Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. *Neural Computation* 4(1):1–58, DOI 10.1162/neco.1992.4.1.1
- Glickman M, Jensen S (2005) Adaptive paired comparison design. *Journal of Statistical Planning and Inference* 127:279–293
- Groot P, Birlutiu A, Heskes T (2010) Bayesian monte carlo for the global optimization of expensive functions. 19th European Conference on Artificial Intelligence (ECAI)
- Guo S, Sanner S (2010) Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. *AISTATS* pp 289–296
- Har-peled S, Roth D, Zimak D (2002) Constraint classification: A new approach to multiclass classification and ranking. In: *Advances in Neural Information Processing Systems 15*, pp 365–379
- Harpale A, Yang Y (2008) Personalized active learning for collaborative filtering. In: *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp 91–98, DOI 10.1145/1390334.1390352
- Herbrich R, Graepel T, Bollmann-Sdorra P, Obermayer K (1998) Learning preference relations for information retrieval
- Heskes T, de Vries B (2005) Incremental utility elicitation for adaptive personalization. In: Verbeeck K, Tuyls K, Nowé A, Manderick B, Kuijpers B (eds) *BNAIC 2005, Proceedings of the Seventeenth Belgium-Netherlands Conference on Artificial Intelligence*, Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Brussels, pp 127–134
- Jin R, Si L (2004) A bayesian approach toward active learning for collaborative filtering. In: *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, AUAI Press, Arlington, Virginia, United States, pp 278–285
- Kanninen B (2002) Optimal design for multinomial choice experiments. *Journal of Marketing Research* 39:307–317
- Lewi J, Butera R, Paninski L (2009) Sequential optimal design of neurophysiology experiments. *Neural Computation* 21(3):619–687
- Lewis D, Gale W (1994) A sequential algorithm for training text classifiers. In: *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., New York, NY, USA, pp 3–12
- MacKay D (1992) Information-based objective functions for active data selection. *Neural Computation* 4:590–604
- Mackay D (2002) *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA
- McCallum A, Nigam K (1998) Employing EM and pool-based active learning for text classification. In: *International Conference on Machine Learning 1998*, Morgan Kaufmann Publishers

-
- Inc., San Francisco, CA, USA, pp 350–358
- Melville P, Mooney R (2004) Diverse ensembles for active learning. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, ACM, New York, NY, USA, p 74, DOI 10.1145/1015330.1015385
- Minka T (2001) A family of approximation methods for approximate Bayesian inference. PhD thesis, MIT
- Schein A, Ungar L (2007) Active learning for logistic regression: an evaluation. *Machine Learning* 68(3):235–265, DOI 10.1007/s10994-007-5019-5
- Seeger M (2008) Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research* 9:759–813
- Settles B, Craven M (2008) An analysis of active learning strategies for sequence labeling tasks. In: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Morristown, NJ, USA, pp 1070–1079
- Seung H, Oppen M, Sompolinsky H (1992) Query by Committee. In: COLT, ACM, New York, NY, USA, pp 287–294
- Yu K, Schwaighofer A, Tresp V, Ma W, Zhang H (2003) Collaborative ensemble learning: Combining collaborative and content-based information filtering. In: In Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence, pp 616–623
- Yu K, Tresp V, Schwaighofer A (2005) Learning Gaussian Processes from Multiple Tasks. In: Proceedings of the 22nd International Conference on Machine Learning
- Yu K, Bi J, Tresp V (2006) Active learning via transductive experimental design. In: International Conference on Machine Learning, New York, pp 1081–1088